PROCEEDINGS *of the* THIRD BERKELEY SYMPOSIUM ON MATHEMATICAL STATISTICS AND PROBABILITY

Held at the Statistical Laboratory
University of California
December, 1954
July and August, 1955

VOLUME I

CONTRIBUTIONS TO THE THEORY OF STATISTICS

EDITED BY JERZY NEYMAN

For further effective support of the Symposium thanks must be given the National Science Foundation, the United States Air Force Research and Development Command, the United States Army Office of Ordnance Research, and the United States Navy Office of Naval Research.

UNIVERSITY OF CALIFORNIA PRESS
BERKELEY AND LOS ANGELES
1956

TWO APPROXIMATIONS TO THE ROBBINS-MONRO PROCESS

J. L. HODGES, JR. AND E. L. LEHMANN*
UNIVERSITY OF CALIFORNIA

1. Introduction

The following process was introduced by Robbins and Monro [1]. For each real number x let $Y(x)$ be a random variable such that $E[Y(x)] = M(x)$ exists. We assume that M is Borel measurable, that the regression equation $M(x) = \alpha$ has a single root θ, which we wish to estimate, and that $(x - \theta)[M(x) - \alpha] > 0$ for all $x \neq \theta$. An initial value x_1 and a sequence $\{a_n\}$ of positive numbers are selected. The $(n + 1)$st approximation to θ is defined inductively by the formula

$$x_{n+1} = x_n - a_n [Y(x_n) - \alpha]. \tag{1.1}$$

In [1], [2], conditions were investigated under which X_n tends to θ in mean square, and in [3], [4] for convergence with probability 1.

The statistician is naturally concerned with the speed of convergence, and with the choice of coefficients $\{a_n\}$ to maximize the speed. This problem was attacked by Chung [5] who studied the asymptotic behavior of the moments of X_n, and thereby was able to prove asymptotic normality under certain conditions.

Chung considers two cases, using different coefficients a_n and getting variances of different orders in the two:

(i) The "quasi-linear" case (theorem 9). Here, $a_n = c/n$, and $\sqrt{n}\ (X_n - \theta)$ tends in law to the normal distribution $N[0, \sigma^2c^2/(2a_1c - 1)]$ where $a_1 = M'(\theta) > 0$ and σ^2 is the variance of $Y(\theta)$. The variance of X_n tends to 0 with the speed $1/n$ which a statistician would hope for. Chung proves optimum properties for these estimates. Among the assumptions of theorem 9 we mention particularly

$$\lim_{|x| \to \infty} \frac{M(x)}{x} > 0, \tag{1.2}$$

which as Chung emphasizes is quite restrictive from the point of view of statistical applications, since it is not satisfied in any problem in which $M(x)$ is bounded. For example, the quantal response problem (in which up-and-down methods generally had their origin and to date their most important applications) is excluded.

(ii) The "bounded case" (theorem 6). Here, $M(x)$ is bounded, but unfortunately the coefficients a_n are taken to be $1/n^{1-\epsilon}$ where ϵ must exceed a positive number $1/2(1 + K_4)$ whose value depends on the problem. Chung now shows $n^{(1-\epsilon)/2}(x_n - \theta)$ to have a normal limit, so that the variance of x_n tends to 0 with the speed $1/n^{1-\epsilon}$. The statistician is naturally unhappy with estimates of such great variability.

* This paper was prepared with the partial support of the Office of Naval Research, and of the Office of Ordnance Research, U.S. Army, under Contract DA-04-200-ORD-355.

A main purpose of this paper is to point out that (1.2) is not an essential condition of Chung's theorem 9. Relying heavily on Chung's analysis, and using a result on convergence with probability one, obtained independently by Blum [3], Kallianpur [4], and Kiefer and Wolfowitz [3], we are in fact able to prove the result of theorem 9 under assumptions about the model somewhat weaker than those of theorem 6. As a consequence, we recommend that the coefficients $1/n^{1-\epsilon}$ not be used in statistical practice.

It should be emphasized that Chung's penetrating analysis of the moments of X_n actually proves more than the mere convergence in law to the normal which is asserted in theorems 6 and 9. Since he knows the asymptotic behavior of the variance, he can assert optimum properties concerned with squared error for his estimates in theorem 9. Our result is in this regard much weaker, since our method involves a truncation that prevents any control over the variance. We discuss in section 3 the statistical significance of the two ways of studying the asymptotic variance of estimates, which are involved here.

An alternative approach to the performance characteristic of the Robbins-Monro estimates is presented in section 5. There, instead of studying the limiting distribution of the actual estimates, we examine the actual variance of the estimates obtained from the linear model approximating to the actual model. For the linear model it is possible to compute the exact variances, and it is comforting to observe that the limiting values of the exact variances of the linear model agree with the variances of the limiting distribution of the actual estimates.

2. The bounded case with harmonic coefficients

Our considerations are based on theorem 9 of Chung [5], which states that $\sqrt{n}$ $(X_n - \theta)$ tends in law to $N[0, \sigma^2c^2/(2a_1c - 1)]$ under the following assumptions:

(I) $a_1 = M'(\theta) > 0$,

(II) for every $\delta > 0$, $\inf_{|x-\theta|>\delta} |M(x) - a| > 0$,

(V) $E[Y(x) - M(x)]^2 = \sigma^2 > 0$ for all x,

(VI) (a) $|M(x)|$ is bounded in every finite interval, (b) $0 < \lim_{|x|\to\infty} [M(x)/x]$, and (c) $\overline{\lim}_{|x|\to\infty} [M(x)/x] < \infty$,

(VII) for every even positive integer p,

$$E[Y(x) - M(x)]^p \leqq K_{21}(p) < \infty,$$

(2.1)

$$a_n = \frac{c}{n} \quad \text{where} \quad c > \frac{1}{2K}$$

and K is any positive number not greater than $\inf [M(x) - a]/(x - \theta)$.

As Chung remarks (footnote 4), (V) may be replaced by the weaker assumption that $V(x) = E[Y(x) - M(x)]^2$ is continuous and has the value σ^2 at $x = \theta$. We observe that the proof of the theorem also remains valid with only minor changes if we merely assume $na_n \to c$ instead of $na_n = c$.

We shall now show that Chung's theorem 9 remains valid if we remove assumption (VIb). This permits the theorem to be applied to problems (such as the bio-assay problem) in which $M(x)$ is bounded.

It was discovered independently by Blum [3], Kallianpur [4], and Kiefer and Wolfo-

witz [3] that X_n tends to θ with probability one under certain conditions. The conditions of Blum's theorem 1 are implied by the assumptions of theorem 9.

Let A be any positive number, and suppose that a model $\mathcal{M}$ satisfies all of the assumptions of theorem 9 except possibly (VIb). Let $K' = \inf_{|x-\theta| \leqq A} [M(x) - a]/(x - \theta)$, and construct a new model $\mathcal{M}'$ by defining $Y(x)$ to have the same distribution as before if $|x - \theta| \leqq A$, but the normal distribution $N[K'(x - \theta), 1]$ otherwise. The new model satisfies the conditions of theorem 9. Now introduce a sequence of coefficients a_n such that $na_n \to c > 1/2K'$, and consider the process $X_1, X_2, \cdots$ generated under the model $\mathcal{M}$. Inasmuch as X_n converges to θ with probability one, we can associate with each $\epsilon > 0$ a number $N(\epsilon)$ such that the probability is at least $1 - \epsilon$ that $|X_n - \theta| < A$ for all $n > N(\epsilon)$.

We define a new process, whose starting point is $X_1' = X_{N+1}$ and which is generated by the model $\mathcal{M}'$ and coefficients $a_n' = a_{n+N}$. We shall have $X_m' = X_{N+m}$ for all m with probability at least $1 - \epsilon$. Let us denote generically the distribution of a random variable Z by F_Z. We observe that for all m $|F_{X_{N+m}} - F_{X_m'}| < \epsilon$. Since the process X' satisfies the conditions of theorem 9, we can choose m so large that $|F_{\sqrt{m}(X_m' - \theta)} - \Psi| < \epsilon$ where Ψ is the normal distribution function with mean 0 and variance $\sigma^2c^2/(2ac - 1)$. Consequently $|F_{\sqrt{m}(X_{N+m} - \theta)} - \Psi| < 2\epsilon$, from which it follows that for m sufficiently large $|F_{\sqrt{N+m}(X_{N+m} - \theta)} - \Psi| < \epsilon$. Since ϵ is arbitrary, our result follows.

The above proof was obtained by the authors in cooperation with Professor Charles Stein, whose help we gratefully acknowledge.

3. Two measures of asymptotic accuracy

Two measures of the limiting accuracy of a sequence of estimates are used (and sometimes confused) in the literature. We shall in this section briefly discuss some of their relationships, with particular reference to our problem.

Consider a sequence of estimates $X_1, X_2, \cdots$ for a parameter θ, and let $Y_1, Y_2, \cdots$ be the corresponding sequence of errors of estimate, appropriately normed. We consider the normed error variances, $v_n = E(Y_n^2)$, which may approach a limit u as n tends to infinity, and say in this case that u is the *asymptotic error variance* of the sequence of estimates. It often happens that Y_n tends in law to a random variable Y (usually normal), and that Y possesses an error variance $w = EY^2$. In the usual situation, $E(Y) = 0$ so that w is the actual variance of Y. We then call w the *asymptotic variance in law* (or *asymptotic normal variance* in the normal case). It is easy to show that $w \leqq u$, but strict inequality is possible.

Both u and w are used in the literature as measures of precision of estimates. The significance of w lies in the approximate probability statements which can be based on it. For example, if $\sqrt{n}\ (X_n - \theta)$ has asymptotic normal variance w, while $\sqrt{n}\ (X_n' - \theta)$ has asymptotic normal variance $w' > w$, then for each $A > 0$, $P[|X_n - \theta| < A/\sqrt{n}] > P[|X_n' - \theta| < A/\sqrt{n}]$ for all sufficiently large n. Thus w is an appropriate measure if we are more interested in the frequency of errors greater than $A/\sqrt{n}$ than in their magnitude. On the other hand, u is an approximation to v_n if n is large, so that it weights large errors much more heavily.

In practice our estimates are usually truncated, which suggests that we consider the random variables Y_n^A obtained by truncating Y_n at $\pm A$. Let v_n^A denote the error variance of this truncated estimate, and call $t^A = \lim_{n \to \infty} v_n^A$, if it exists, the asymptotic error vari-

ance truncated at A. It is easy to show that $t^A \to w$ as $A \to \infty$, which suggests that w has the interpretation of a limiting truncated error variance. This notion does not require that we introduce the limit law or even that a limit law exist. We note that more precisely $\lim_{A\to\infty} \lim_{n\to\infty} v_n^A = \liminf_{A,n\to\infty} v_n^A \leqq \limsup_{A,n\to\infty} v_n^A = \lim_{n\to\infty} \lim_{A\to\infty} v_n^A$, provided the limits involved all exist. This is a simple consequence of the fact that v_n^A is for each n a nondecreasing function of A.

Since in practice both n and A are finite, it is not clear whether w or u (in those cases in which $w < u$) will be the better approximation to v_n^A. However, if $u > w$, this can only mean that very large errors occur with very small probability. The situation is similar to that in the Petersburg paradox. The usual human practice of not attaching undue importance to large errors which are extremely unlikely to happen would lead to the use of w in preference to u. Another argument which also supports this choice lies in questioning the reasonableness of squared error as a loss function when the errors are very large.

As a consequence of these considerations, we are inclined to use the asymptotic normal variance as a reasonable means of appraising the estimates discussed in section 2 when n is large. In particular, we recommend that coefficients $a_n \sim c/n$ be used in the quantal response problem. Further, we suggest that c be chosen so as to minimize $\sigma^2c^2/(2a_1c - 1)$. This leads to $c = 1/a_1$ and reduces the asymptotic normal variance to σ^2/a_1^2. (In practice, of course, it will usually be necessary to guess at the value of a_1.)

It is hardly necessary to remark that the only interest in any asymptotic theory resides in the hope that it will provide a useful approximation for the values of n with which we are dealing. Thus, for example, we use the asymptotic normal variance as an approximation to the variance of a normal distribution which approximates to the actual distribution of the estimate. For many statistical problems the only way of appraising the accuracy of these approximations lies in comparing them with computed values for small n or sampling experiments with moderate n. In the next section we consider computed values of the variance for small n of an approximate model, leading to conclusions about the choice of c in general agreement with those given above.

4. A linear approximation

If one attempts to apply the asymptotic normal theory of section 2, two difficulties arise. As with most asymptotic theories, it is not known how large n must be before the theory becomes applicable. Furthermore, the theory holds only if $c > 1/2K$, where K is any positive number satisfying $K \leqq \inf [M(x) - \alpha]/(X - \theta)$. An examination of the proof shows that in this condition the infimum may be restricted to the values $|x - \theta| \leqq A$, where A is an arbitrarily small positive number. Since we assume $M'(\theta) = a_1 > 0$, it is therefore enough to require $c > 1/2a_1$. This is consistent with the recommendation made in section 3 that $c = 1/a_1$. In practice, however, a_1 is usually not exactly known, and one might be tempted to use a "safe" small *a priori* estimate for a_1, and a correspondingly large c, to avoid the possibility that $c \leqq 1/2a_1$ in which case the estimates have unknown behavior. This tendency would produce a bias towards values of c too high for greatest efficiency, but as $c^2/(2c - 1)$ increases slowly when c increases beyond 1, it would be natural to prefer a c which might be too large to one which might be too small.

We shall now present an alternative approach which (while it also has drawbacks) does work for all values of $c > 0$ and does provide measures of precision for finite n. The

current approach is based on replacing the actual model by a simpler linear model, for which the actual error variances can be computed. Specifically, we assume $M(x) = \alpha + \beta(x - \theta)$ and $V(x) = \tau^2$, where β and τ^2 are known constants. We might take $\beta = a_1$, $\tau^2 = \sigma^2[= V(\theta)]$, obtaining a model which is a good approximation to the actual one when x is near θ; alternatively, we might attempt to fit a straight line to that portion of $M(x)$ where the x_n are likely to fall. To simplify the notation, we shall set $\alpha = \theta = 0$.

It is easily shown that

$$E(X_{n+1}^2) = (1 - \beta a_n)^2 E(X_n^2) + a_n^2 \tau^2 \tag{4.1}$$

from which it follows that $E(X_{n+1}^2)$ equals

$$E(X_1^2)\left[\prod_{\nu=1}^{n}(1 - \beta a_\nu)\right]^2 + \left(\frac{\tau}{\beta}\right)^2 \sum_{k=1}^{n} (\beta a_k)^2 \prod_{\nu=k+1}^{n} (1 - \beta a_\nu)^2. \tag{4.2}$$

Both of the terms of (4.2) have a significance. Since $X_{n+1} = X_n - a_n Y_n$, $E(X_{n+1}) = (1 - \beta a_n)E(X_n)$, so that given $X_1 = x_1$, $E(X_{n+1}) = x_1 \prod_{\nu=1}^{n} (1 - \beta a_n)$. If we square and take expectations, we find that the first term of (4.2) is the expected squared bias of the estimate. It is the contribution to the total error variance of the error of our initial guess x_1, and vanishes if $x_1 = \theta = 0$. The second term of (4.2) is independent of X_1, and represents the variance component of the error variance.

We shall now specialize to harmonic coefficients $a_n = c/n$, so that (4.2) becomes

$$E(X_1^2)\,\varphi_n(c\beta) + \left(\frac{\tau}{\beta}\right)^2 \psi_n(c\beta) \tag{4.3}$$

where

$$\varphi_n(c) = \prod_{\nu=1}^{n}\left(1 - \frac{c}{\nu}\right)^2, \quad \psi_n(c) = \sum_{k=1}^{n}\left(\frac{c}{k}\right)^2 \prod_{\nu=k+1}^{n}\left(1 - \frac{c}{\nu}\right)^2. \tag{4.4}$$

It is easily seen that $\varphi_n(c) = 0$ if $c \leqq n$ is a positive integer, and that for nonintegral $c > 0$, $\varphi_n(c)$ is of the order n^{-2c}. The analysis of the second term is more complicated, but it can be shown that it is asymptotically equivalent to

$$\frac{\tau^2 c^2}{n(2c\beta - 1)} \quad \text{if} \quad c > \frac{1}{2\beta} \tag{4.5}$$

$$\frac{\tau^2 \log n}{4n} \cdot \frac{1}{\beta^2} \quad \text{if} \quad c = \frac{1}{2\beta} \tag{4.6}$$

$$\frac{R_0/\Gamma^2(1 - c\beta) + \tau^2 c^2 \rho(c\beta)}{n^{2c\beta}} \quad \text{if} \quad 0 < c < \frac{1}{2\beta} \tag{4.7}$$

where

$$\rho(c) = \sum_{k=1}^{\infty} \frac{1}{k^2} \frac{1}{\Gamma^2(1 - c) \prod_{\nu=1}^{k}\left(1 - \frac{c}{\nu}\right)^2}. \tag{4.8}$$

These formulas may be compared with the formulas (31), (33), and (35) of Schmetterer [9], who obtained the same asymptotic orders for an upper bound on $E(X_n^2)$ without assuming linearity.

We observe that (4.5) becomes the familiar formula $\sigma^2/n(2ca_1 - 1)$ if we take $\beta = a_1$ and $\tau^2 = \sigma^2$; that is, the asymptotic variance of the linear model which replaces $M(x)$ by its tangent at θ coincides with the asymptotic normal variance of the original model. However, our main interest here is in actual values of the expressions (4.4) and not in asymptotic theory, and in practice one might choose a β different from a_1 to obtain a better fit to $M(x)$ for small or moderate n.

Tables I and II present a few values of $n\varphi_n(c)$ and $n\psi_n(c)$which facilitate the computation of (4.3). We note that $n\varphi_n(1) \equiv 0$, while $n\varphi_n(1/2) \to 1/\pi = 0.318 \cdots$ by virtue of Wallis' product. For values of n larger than 30, one may use the approximation $\varphi_n(c) = \varphi_{30}(c) \cdot (30.5/n + 1/2)^{2c}$.

TABLE I

$n\varphi_n(c)$

n	$c=$				
	0.2	0.4	0.6	0.8	1.2
5	1.878	.593	.140	.017	.003
10	2.878	.698	.125	.012	.001
15	3.707	.763	.116	.009	.001
20	4.417	.811	.110	.008	.000
25	5.057	.850	.106	.007	.000
30	5.648	.883	.102	.006	.000

The recursion formula

$$\psi_n(c) = \left(\frac{c}{n}\right)^2 + \left(\frac{1-c}{n}\right)^2 \psi_{n-1}(c) \tag{4.9}$$

permits easy computation of $\psi_n(c)$. For values of $\psi_n(c)$ not in the table one may use the quick approximation

$$\frac{n^3\psi_n(c)}{c^2} \doteq \frac{(n-1)(n-2)}{2c-1} + 2(2-c)^2(n-1) + n \tag{4.10}$$

which is based on quadratic interpolation of $n\psi_n(c)$ against $1/n$ at the values $1/n = 0$, $1/2$, 1.

TABLE II

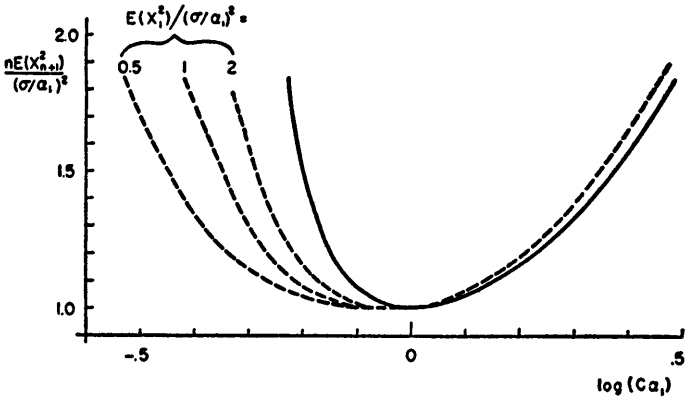
$n\psi_n(c)$

n	$c=$							
	0.2	0.4	0.6	0.8	1.2	1.6	2.0	3.0
5	0.192	0.502	0.748	0.904	1.075	1.253	1.500	2.300
10	0.326	0.700	0.889	0.961	1.048	1.203	1.407	2.008
15	0.435	0.830	0.963	0.985	1.041	1.189	1.381	1.932
20	0.530	0.929	1.011	0.998	1.037	1.182	1.368	1.896
25	0.616	1.009	1.046	1.007	1.035	1.178	1.361	1.877
30	0.696	1.077	1.074	1.013	1.034	1.176	1.356	1.867
∞	∞	∞	1.800	1.067	1.029	1.164	1.333	1.800

Note the good agreement between the values at $n = 30$ and $n = \infty$ except for c near or below $1/2$.

We now examine the choice of c from the point of view of the linear approximation,

taking $\beta = a_1$, $\sigma^2 = V(\theta)$. The value of $E(X_{n+1}^2)$, as given by (4.3), depends on $E(X_1^2)$ as well as on the quantities n, $(\sigma/a_1)^2$ and ca_1 which enter into the determination of the asymptotic normal variance. (This is an advantage of the linear theory, since in practice the accuracy of the initial guess will be important.) The figure shows $nE(X_{n+1}^2)/(\sigma/a_1)^2$



for $n = 20$, and for $E(X_1^2)/(\sigma/a_1)^2 = 1/2$, 1 and 2, as functions of log (ca_1); such charts are quickly sketched with the aid of the tables. For comparison, we also show as a solid line $(ca_1)^2/(2ca_1 - 1)$, which corresponds to $n = \infty$, $E(X_1^2)$ arbitrary; or to the asymptotic theory of section 2.

An examination of the chart suggests that in general the two theories lead to similar conclusions as to the choice of c and n in that both suggest $c = 1/a_1$ as a good value, and for this value lead to about the same n for given variance. There are however differences. The best agreement occurs for ca_1 above 1, the worst for ca_1 near or below 1/2. We have here an example in which an asymptotic theory is somewhat misleading. When $ca_1 = 1/2$, the asymptotic normal variance is infinite; the linear variance tends to ∞ as $n \to \infty$, but only slowly. In general, the linear analysis leads to the choice of a smaller c than the asymptotic theory, particularly if $E(X_1^2)/(\sigma/a_1)^2$ is small. The effects noticed vary inversely with the size of n, the linear model tending to agree with the asymptotic results as $n \to \infty$.

The main drawback of the linear model is, of course, the fact that we do not know how nearly linear $M(x)$ must be, nor how nearly constant $V(x)$ must be, in order that the linear approximation will represent what actually happens. The only evidence on this point known to us consists in a sampling experiment [8]. There it is found that the linear theory is in reasonable agreement with the data, although intuitively the model is quite "nonlinear." Further experience is needed on this question.

5. Parametric estimation

The greatest advantage of the Robbins-Monro scheme is the fact that it provides consistent estimation in a broad nonparametric situation. However, it may also be applied to parametric estimation problems. Hitherto we have supposed that the distributions of $Y(x)$ were arbitrary except for some restrictions on the first two moments. In many problems one can however assume that the distributions of $Y(x)$ are known except for the value of a real parameter γ. The parametrization (which of course is not unique) will

be chosen so that γ is the quantity which is to be estimated. We shall use the notation $E[Y(x)] = M_\gamma(x)$, $V[Y(x)] = V_\gamma(x)$, and still assume that these functions satisfy the conditions of the theorem of section 2. Since γ now determines the model, θ is a function of α and γ, and we shall make the additional assumption that for each α, θ is a $1 - 1$ function of γ, permitting us to invert to find $\gamma = h_\alpha(\theta)$. We may then use the Robbins-Monro estimates X_n for θ to provide estimates $h_\alpha(X_n)$ for γ. Our problem now becomes that of choosing a_n (and perhaps also α) to minimize the asymptotic normal variance of these new estimates.

It is then obvious that the estimates $h_\alpha(X_n)$ are such that $\sqrt{n}\,[h_\alpha(X_n) - \gamma]$ is asymptotically normal with mean 0 and variance

$$\frac{[h'_\alpha(\theta)]^2\sigma^2 c^2}{2a_1 c - 1}. \tag{5.1}$$

As illustration, consider the quantal response problem, in which $Y(x)$ is capable of assuming only the values 0 and 1, so that $M_\gamma(x) = P[Y(x) = 1]$ and $V_\gamma(x) = M_\gamma(x)[1 - M_\gamma(x)]$. We take $0 < \alpha < 1$, and estimate θ by means of a Robbins-Monro scheme. This provides a sequence of estimates X_n such that $\sqrt{n}\,(X_n - \theta) \sim N[0, V(\theta)c^2/(2a_1c - 1)]$.

Let the partial derivatives of $M_\gamma(x)$ with respect to x and to γ be denoted respectively by $M'_\gamma(x)$ and $M^*_\gamma(x)$. Given α, the best value of $c = \lim na_n$ is $c = 1/M'_\gamma(\theta)$; the resulting asymptotic normal variance is $\sigma^2/[M'(\theta)]^2$.

In parametric problems such as the present one, we may be able to choose α with an eye to minimizing (5.1), which becomes

$$\frac{[h'_\alpha(\theta)]^2\sigma^2(\theta)}{[M'(\theta)]^2} \tag{5.2}$$

where we now make explicit the dependence of σ^2 on θ.

On differentiating the identity $M_{h_\alpha(\theta)}(\theta) = \alpha$ with respect to θ, we get $h'_\alpha(\theta) = -M'_\gamma(\theta)/M^*_\gamma(\theta)$ and therefore (5.2) becomes

$$\frac{M_\gamma(\theta)[1 - M_\gamma(\theta)]}{[M^*_\gamma(\theta)]^2}. \tag{5.3}$$

The numerator of (5.3) being equal to $\alpha(1 - \alpha)$, it is seen that the value of α that minimizes (5.3) will be independent of the unknown γ provided $M^*_\gamma(\theta)$ factors into a function of γ alone and a function of α alone. This is the case in particular if $M_\gamma(x)$ is a function of $x - \gamma$ or $x\gamma$ or more generally if there exist functions r, s and t such that $M_\gamma(x) = r[s(\gamma)t(x)]$.

As an illustration we consider the bio-assay form of the quantal response problem in which it is customary to take $M_\gamma(x)$ to be a distribution function with γ a location parameter. (Our theory is essentially uniparametric; we assume that the scale parameter of the distribution is known.)

If F is any cumulative distribution function and $0 < \beta < 1$, we may obtain a parametric family by defining $P[Y(x) = 1] = M_\gamma(x) = F[x - \gamma + F^{-1}(\beta)]$. Then γ has the significance of the value of the stimulus x for which probability of response is β;

that is, γ is the "lethal dose 100 β." Formula (5.3) for the asymptotic normal variance now becomes

$$\frac{\alpha(1-\alpha)}{\{F'[F^{-1}(\alpha)]\}^2}. \tag{5.4}$$

It is not surprising that expression (5.4) is independent of β, since the problem of estimating a location parameter remains substantially unaltered when the origin is changed. It happens that (5.4) is minimized by taking $\alpha = 1/2$ when F is either normal or logistic, which are the most common choices for F.

The estimation of γ = lethal dose 100 β does not require the use of a parametric model, since we may set α equal to β, with $\gamma = \theta$, and thus estimate γ directly by X_n as in section 2. The advantage of this approach is that it requires very little assumption about the form of F; its disadvantage is that there may be a substantial loss of efficiency, particularly if β is not near 1/2.

Further illustrations are provided by the testing problems treated in [6]. In both of the specific situations analyzed there, both x and γ are essentially nonnegative and $M_\gamma(x)$ is a function of the product $x\gamma$ only, say $M_\gamma(x) = m(x\gamma)$. It is easily seen that the restriction of range causes no difficulty.

As an example, consider the problem of estimating the mean bacterial density γ of a liquid by the dilution method. That is, we take a volume x of the liquid at random, and determine whether there is one or more bacteria in it, indicating this event by $Y(x)$. Then $M_\gamma(x) = 1 - \exp(-\gamma x)$ under the usual Poisson assumption. It is easily seen that (5.3) becomes $[\alpha/(1-\alpha)\log^2(1-\alpha)]\gamma^2$, and, whatever be γ, this is minimized by minimizing the first factor. The same extremum problem occurred in [6], in connection with choosing α for maximum asymptotic power. The best α [see equation (19) in [6]] is the root of $2\alpha = -\log(1-\alpha)$, or $\alpha = 0.797$. Thus the following procedure is recommended: use the Robbins-Monro method with $\alpha = 0.797$, and $a_n = 4.92/\hat{\gamma}n$ where $\hat{\gamma}$ is our best *a priori* guess for γ. Our estimate for γ after n steps is $2\alpha/X_{n+1} = 1.594/X_{n+1}$. The asymptotic normal variance is $\gamma^2/4\alpha(1-\alpha) = 1.544\,\gamma^2$.

Actually, the entire family of testing problems considered in [6] can be thus treated as estimation problems; the task of minimizing (5.3) is identical with that of maximizing equation (12) of [6]. A further example there considered relates to the quality control of variability.

We have given a detailed discussion of the estimation of γ in the binomial case since the only interesting examples that we know are of this type. However, we shall sketch briefly an extension of these results to the case of arbitrary distributions belonging to the Darmois-Koopman-Pitman family.

Following the notation of Girshick and Savage [7] let us assume that the generalized density of $Y(x)$ with respect to a measure ψ_x is

$$\frac{1}{\omega(\tau)}\, e^{\tau y} \tag{5.5}$$

where $\tau = \tau(\gamma, x)$. We denote the mean and variance of $Y(x)$ by $M_\gamma(x)$ and $V_\gamma(x)$ respectively and we then have from [7] that

$$\begin{aligned} M_\gamma(x) &= \frac{d}{d\tau}\log\omega[\tau(\gamma, x)] \\ V_\gamma(x) &= \frac{d^2}{d\tau^2}\log\omega[\tau(\gamma, x)]. \end{aligned} \tag{5.6}$$

As usual we define θ by $M_\gamma(\theta) = a$, and solving this equation for γ obtain $\gamma = h_a(\theta)$. Setting $\hat{\gamma}_n = h_a(X_{n+1})$ we then obtain as before the asymptotic normal variance,

$$\frac{[h_a'(\theta)]^2 V_\gamma(\theta)}{[M_\gamma'(\theta)]^2} = \frac{V_\gamma(\theta)}{[M_\gamma^*(\theta)]^2} \tag{5.7}$$

for $\sqrt{n}\,(\hat{\gamma}_n - \gamma)$.

REFERENCES

[1] HERBERT ROBBINS and SUTTON MONRO, "A stochastic approximation method," *Annals of Math. Stat.*, Vol. 22 (1951), pp. 400–407.

[2] J. WOLFOWITZ, "On the stochastic approximation method of Robbins and Monro," *Annals of Math. Stat.*, Vol. 23 (1952), pp. 457–461.

[3] JULIUS R. BLUM, "Approximation methods which converge with probability one," *Annals of Math. Stat.*, Vol. 25 (1954), pp. 382–386.

[4] GOPINATH KALLIANPUR, "A note on the Robbins-Monro stochastic approximation method," *Annals of Math. Stat.*, Vol. 25 (1954), pp. 386–388.

[5] K. L. CHUNG, "On a stochastic approximation method," *Annals of Math. Stat.*, Vol. 25 (1954), pp. 463–483.

[6] J. L. HODGES, JR., "The choice of inspection stringency in acceptance sampling by attributes," *Univ. of California Publ. in Stat.*, Vol. 1, No. 1 (1949), pp. 1–14.

[7] M. A. GIRSHICK and L. J. SAVAGE, "Bayes and minimax estimates for quadratic loss functions," *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley and Los Angeles, University of California Press, 1951, pp. 53–73.

[8] DAN TEICHROEW, "An empirical investigation of the stochastic approximation method of Robbins and Monro," unpublished.

[9] L. SCHMETTERER, "Bemerkungen zum Verfahren der stochastischen Iteration," *Österreich. Ingenieur-Archiv*, Vol. 7 (1953), pp. 111–117.